Data and Metadata. 2024; 3:.349
doi: 10.56294/dm2024.349

ORIGINAL

An Effective Topic Modeling Strategies for Recommender Systems in Crowdfunding Platforms

Un tema eficaz que modela estrategias para sistemas de recomendación en plataformas de financiación colectiva

Suresh Subramanian[1]

[1]College of Information Technology, Ahlia University. Kingdom of Bahrain.

Cite as: Subramanian S. An Effective Topic Modeling Strategies for Recommender Systems in Crowdfunding Platforms. Data and Metadata. 2024; 3:.349. https://doi.org/10.56294/dm2024.349

Submitted: 15-02-2024 **Revised:** 25-05-2024 **Accepted:** 26-09-2024 **Published:** 27-09-2024

Editor: Adrián Alejandro Vitón Castillo

ABSTRACT

Capitalists come up with creative and innovative concepts, but a lack of finance limits their untapped economic potential. There are several channels that new entrepreneurs may use and take advantage of to attract money and other financial resources when beginning a firm thanks to current technology, which has drastically altered the way business is done on a broad scale. An entrepreneur uses the Internet to promote his concept to potential backers through crowdfunding. Online crowdfunding has labored to develop several advanced platforms that may serve as an interface to the fundraising process for a certain concept or project. Typically, the owner of the concept explores the market and does extensive research through a variety of channels, with the Internet assisting in moving ahead and making the idea actual. In truth, the owner of the concept frequently suffers obstacles and financial issues, therefore crowdsourcing helps to alleviate these issues. In this study, machine learning methods were used to train the system on the given data, beginning with the theme, followed by the blurb, which is the topic description, and finally by the topic category. Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA) were employed as machine learning approaches to accomplish the goal. This study employs a variety of text classification algorithms, including Support Vector Machine (SVM), EXtreme Gradient Boosting (XG), K-Nearest Neighbours (KNN), and Random Forest (RF), to propose and forecast subject categories. Each algorithm performed differently in terms of precision, predictability, positive rate, and model correctness. SVM was the highest performance measurement.

Keywords: Latent Dirichlet Allocation; Latent Semantic Analysis; Support Vector Machine; Extreme Gradient Boosting; Random Forest; Crowdfunding.

RESUMEN

A los capitalistas se les ocurren conceptos creativos e innovadores, pero la falta de financiación limita su potencial económico sin explotar. Hay varios canales que los nuevos emprendedores pueden utilizar y aprovechar para atraer dinero y otros recursos financieros al iniciar una empresa gracias a la tecnología actual, que ha alterado drásticamente la forma en que se hacen negocios a gran escala. Un emprendedor utiliza Internet para promover su concepto entre posibles patrocinadores a través del crowdfunding. El crowdfunding en línea ha trabajado para desarrollar varias plataformas avanzadas que puedan servir como interfaz para el proceso de recaudación de fondos para un determinado concepto o proyecto. Por lo general, el propietario del concepto explora el mercado y realiza una investigación exhaustiva a través de una variedad de canales, con Internet ayudando a avanzar y hacer realidad la idea. En verdad, el dueño del concepto frecuentemente sufre obstáculos y problemas financieros, por lo que el crowdsourcing ayuda a aliviar estos problemas. En este estudio, se utilizaron métodos de aprendizaje automático para entrenar el sistema con los datos proporcionados, comenzando con el tema, seguido por la propaganda, que es la descripción del tema y finalmente por la categoría del tema. Se emplearon la asignación latente de Dirichlet (LDA) y el análisis semántico latente (LSA) como enfoques de aprendizaje automático para lograr el objetivo. Este estudio emplea una variedad de algoritmos de clasificación de texto, incluidos Support Vector Machine (SVM), EXtreme Gradient Boosting (XG), K-Nearest Neighbors (KNN) y Random Forest (RF), para proponer y pronosticar categorías de temas. Cada algoritmo funcionó de manera diferente en términos de precisión, previsibilidad, tasa positiva y corrección del modelo. SVM fue la medida de rendimiento más alta.

Palabras clave: Asignación de Dirichlet Latente; Análisis Semántico Latente; Máquina de Vectores de Soporte; Aumento de Gradiente Extremo; Bosque Aleatorio; Financiación Colectiva.

INTRODUCTION

Traditionally, if an entrepreneur needs to raise money to begin a business, assistance will be required and the owner needs to change his business plan strategy, market study, and monetary activity plan, and afterward market his plan to close people or foundations. These kinds of financing sources include banks and private bankers, truly restricting your alternatives to a couple of key partners. Numerous limitations apply to who can support a start-up, new business, and SMEs and the amount they are permitted to contribute, and these guidelines should protect the investors from distributing enormous measures of their spending plans in danger. Since so many of them fizzle, their investors may confront a high danger of losing their capital. Crowdfunding has provided the chance for entrepreneurs to collect an amount of money from anybody to contribute. Crowdfunding outfits a social occasion for anyone with an arrangement to contribute in front of holding up investors.

Financial backers can choose from many ventures and contribute small sums. From taking advantage of a more extensive investor pool to profit more adaptable raising support alternatives, crowdfunding has several advantages such as: using a crowdfunding platform and approaching many approved investors who can observe, deal with, and contribute to fundraising campaigns. By creating a crowdfunding campaign. By marketing or promoting the campaign through online media, email, newsletters, and other marketing strategies. This research will recommend topics for people who are interested in starting their new crowdfunding projects. In this regard, topic modeling methodology will be used to recommend the list of topics using the online crowdfunding project resources available freely. Because of new computing technologies, machine learning now is not like ML in the past. It was used for pattern recognition and the hypothesis that PCs can perceive without being programmed to accomplish specific missions; academics keen on AI wanted to see if PCs could perceive data. The iterative piece of ML is a huge fact that as models are acquainted with new data, they can straightforwardly change. They acquire from past computations to make trustworthy, repeatable decisions and results. Most businesses working with a great deal of data have seen the assessment of ML advancement. By social occasion encounters from this data – consistently persistently – affiliations can work even more beneficially or gain a good situation over competitors.

ML algorithms accomplished new predominance as AI was engaged in indisputable quality.[23,24] Profound learning models execute the most excellent AI applications. ML platforms are among enormous business progression's most veritable territories, with most basic vendors, including Amazon, Google, Microsoft, IBM, and others, rushing to initial clients for phase privileges that comprehend the extent of ML exercises, containing data assortment, data arrangement, data characterization, model structure, planning and application sending. Topic modeling is a machine learning technique that can parse a huge number of documents, presentations, and other "text" materials provided by start-ups and SMEs, detecting words and phrase patterns within them. Also, naturally clustering word gatherings and comparative expressions that best describe these documents. This is known as 'unsupervised' ML since it does not need a predefined rundown of tags, keywords, or preparing data that has been recently classified by people. Over the long years, Topics in a document corpus advance, TM without considering time will perplex topic disclosure. When TM considers the time is called topic evolution. Topic modeling can uncover significant shrouded data in the document corpus, permitting distinguishing topics with the presence of time, and examining their advancement within the time frame. There are great areas that can utilize topic evolution models. A regular model would be this way: a specialist needs to pick a research subject in a specific field and might want to realize how this topic has developed over the long run and attempt to distinguish those documents that clarified the topic.

In figure 1 topic modeling engages counting words and grouping similar patterns to specific topics within unstructured data. As an example, airline companies want to know what customers are planning for summer vacations and where they want to visit. Rather than going through hours experiencing stacks of feedback and rush call centers, the companies could analyze the regional customers' feedback and posts on social media and analyze their content through a topic modeling algorithm. It is extremely helpful for the aim of document clustering, arranging huge blocks of textual data, information retrieval from unstructured documents, and feature selection. Diverse researchers are utilizing TM for recruitment industries and news agencies where they intend to bring out latent highlights of sets of duties and guide them to the right candidates. They are used to master huge datasets of messages, client feedback, and client web-based media profiles.

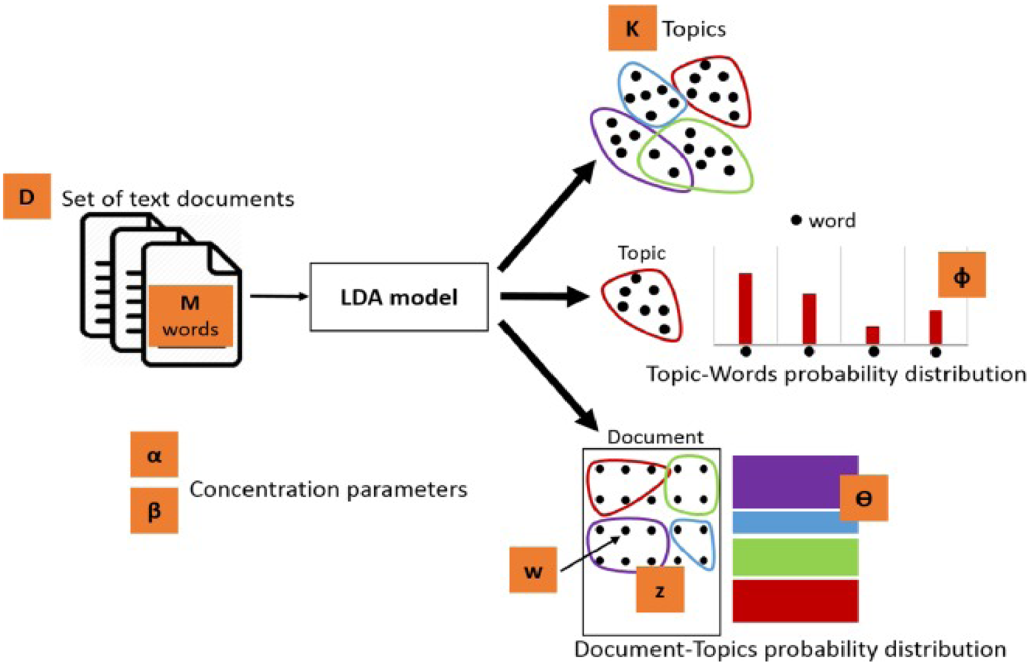

Figure 1. Text model using LDA model

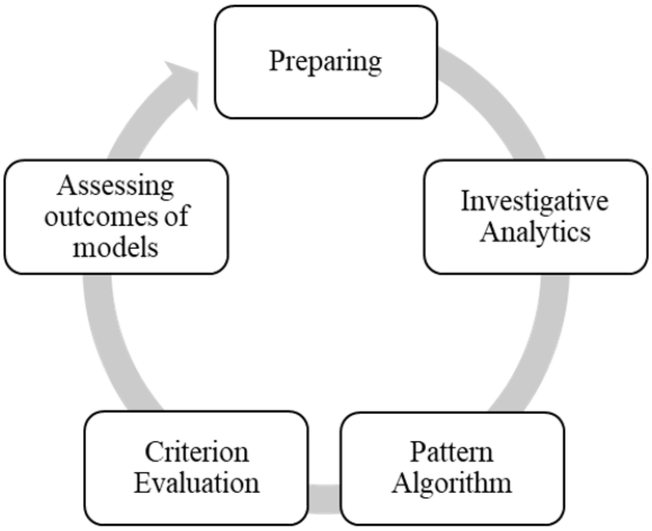

Figure 2. Analysis and design of text architecture

With the rapid explosion of the electronic documents archive in crowdfunding the need to have a superior method to deal with it has been increased. It requires utilizing new strategies or instruments that manage automatically arranging, searching, ordering, and perusing enormous collections. Figure 2 gives a clear idea about text architecture analysis and design. The premise of the present research of Machine Learning (ML), has grown new models to discover the pattern of the text in archive assortments. These models are called "Topic modeling" (TM). It's characterized by the ability to investigate a large unclassified text and find the pattern of the cluster word that is usually repeated and joined together to build the structure of the documents. Furthermore, it can associate words with comparable meanings and recognize the use of words depending on various meanings. The concept of TM is terminology that can deal with documents and these documents are a combination of topics that are distributed. TM is considered a reproductive model for documents. As well as it determines the basic probabilistic method to classify the documents. Topic modeling representation is shown in figure 3.

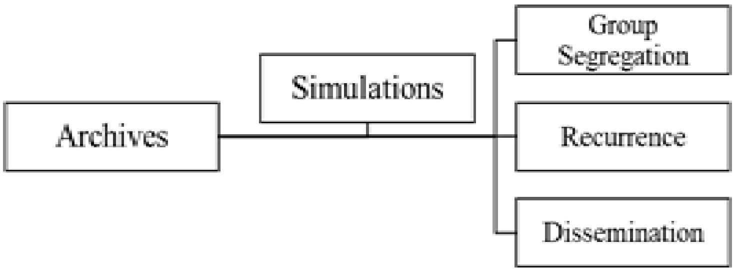

Figure 3. Text segregation

Jin W[1] researched machine learning and its algorithms and development. The researcher divided the ML into three types: supervised, unsupervised, and reinforcement learning. Supervised learning has a place with a generally fundamental learning strategy. This learning technique alludes to the foundation of comparing learning objectives by individuals before learning. During the underlying preparation of the machine, the machine depends on data technology to become familiar with the requirements of learning. To gather fundamental data d, we should slowly complete the necessary learning content in a regulated climate. The so-called unsupervised learning implies that the machine doesn't specify the substance in a specific way during the whole learning procedure yet depends on the actual machine to finish the analysis of information data. Practically, the operation strategy is to allow the machine to become familiar with the fundamental ideas and substance, and afterward give the machine enough opportunity to finish a progression of content learning, including concepts and substance like the essential standards. In the last type which is the reinforcement learning in a particular application procedure, the data gathered in the past period will be utilized. It coordinates and cycles the input data of a specific part to shape a closed loop of information processing. Kolluri J et al[2] insisted that ML approaches can subsequently construct a model using the typical training data. ML algorithms fix input occurrences of normal data and get a model. That is then set up to gather noticed information thusly as either normal or abnormal. ML models can regularly manage little changes or varieties in the noticed data. It can get the inner independence conditions or the more significant relationship between the proportions of the noticed data.

Moreover, in the training stage, it requires a lot of data. Regardless, when the model is conveyed ensuing examination is customarily, and beneficial. Research about text mining, and they went through Machine Learning procedures and typologies. They found out that unsupervised machine-learning techniques are used to categorize text that has an invisible structure in texts for which no predefined categorization issues. It utilizes the features of the texts to practically explore the latent categories and then specify units of texts to those categories. Unsupervised text-categorization methods are characterized by some features. First, they require only a rare human involvement in the pre-analysis and analysis phases. Second, they create reproducible output since there is no human interaction. Third, the unsupervised technique demands a focused post-analysis stage that is time-absorbing to the researchers.

Standard LDA algorithms are much more efficient when choosing Frequency Domain (FD) features in their research. Classification tasks were successfully employed by Artificial Neural Network.[3] The statistical learning techniques comprise the foundation of intelligent software that is utilized to create machine learning. Since ML algorithms expect data to be trained on, the order should have an association with any data set. Knowledge Discovery from Data (KDD) and data mining are examples of machine learning. Naqa I.E et al.[4] believed that ML is a developing part of the topological method which is intended to imitate human intelligence by gaining knowledge from the encompassing surroundings. It looks like a working pony in the current generation and that's why it's so-called big data. Strategies depend on ML employed effectively in different sectors going from pattern recognition, PC vision, space apparatus designing, money, diversion, and computational science to biomedical and clinical utilizations. Hannigan T. R.[5] revealed that Topic modeling has created accuracy and empowered further experiences in studies of novelty and knowledge elements, in this way encouraging the age of new approaches in an assortment of development-related contexts. TM gives significant focal points over conventional techniques, for example, checks of patent filings or resulting references, finding the available classification techniques that weren't intended to catch novel and rising concepts. By straightforwardly utilizing the intellectual context of the text, TM expands the customary evaluation of effect in information fields.

Moreover, by isolating evaluations of effect from those of information itself, TM has progressed the approach by enabling academics to create more exact intends to experimentally test contending hypothetical mechanisms. These deployments of TM may assist researchers specify continuing inquiries in the administration literature by determining the novelty job with institutional rationales or depicting the role of advancement and limits with standards. Hossain M et al.[6] indicated a study about TM and they found that topic modeling is a method that utilizes statistical associations of words in a text to produce latent topics—clusters of co-occurring words

https://doi.org/10.56294/dm2024.349

that collectively show higher sequence ideas—but short of help of predefined, unequivocal word references or interpretive limitations. Also, they believe that creating topics utilizing statistical probabilities has three key advantages. First, to begin with, analysts do not have to implement word references and interpretive data standards. Second, the methods empower the significant subjects' definitions that individual readers cannot detect. Third, it is considered polysemy because the topic is not fundamentally unrelated; singular words show up across the topic with varying probabilities, and the words themselves may overlap or cluster. Shafqat, W et al.[7] reported that crowdfunding has risen to be a cutting-edge and provocative procedure to pull in a significant number of investors to imaginative new startups. The possibility that anybody with an imaginative idea can begin to gather funds for an item, and the challenging procedure to get to enterprise authenticity, develops a feeling of caution for the crowdfunding platforms. A desperate need to establish a proposal framework to recommend the investors as per their interests. It is fundamental to have a sensible measure of data for these proposal frameworks to perform well in the research technique of machine learning, text classification, topic modeling, and sentiment analysis. Topic Identification, Sentiment Analysis, Language recognition, and intent detection are the most well-known uses of text classification.[10] Topics extractions using Mult multi-objective genetic algorithm (MOGA) are not done by many researchers. This research applied MOGA to cluster the texts to extract the topmost 5 words from each cluster.[11]

Table 1. Comparison of various techniques

No	Research Title	Existing methods	Strengths
1	Research on Machine Learning and Its Algorithms and Development[1]	The research explained the algorithms theoretically without going deep into practical deployment.	The research explained the fundamental information about ML in a simple way that is easy to understand
2	Text classification using Machine Learning and Deep Learning Models[2]	Did not include the basic implementation of these classifiers	The research includes the different machine learning typologies that are divided into the text classifiers that belong to each typology
3	Topic Modelling and Opinion Analysis On Climate Change Twitter Data Using LDA And BERT Model[21]	Not related to crowdfunding & related topic collection	The research applied the LDA to list the various topics BERT model was applied to classify the sentiments
4	Text mining for information systems researchers: an annotated topic modeling tutorial[8,9]	The research did not mention how to overcome these challenges.	The Researchers declared the challenges that may occur while applying TM techniques
5	Classification of EMG Signals: using DWT Features and ANN Classifier[3]	LDA, ANN	FDs outperform in sequential signal processing on LDA.
6	Machine Learning in Radiation Oncology[4]	The research did not focus on the algorithms that are applied in the field	The application of Machine Learning in biomedicine
7	Topic modeling in management research: rendering new theory from textual data[5]	Text classification methods such as SVM, RNN, KNN, etc were not included in the study	Satisfying explanation about Topic Modelling Rendering in Theory-Building Spaces
8	Crowdfunding: motives, definitions, typology and ethical challenges[6]	Statistics about crowdfunding did not include	Defining the common terminology in crowdfunding. Combining various crowdfunding typologies from different aspects.
9	Multi-label Classification for Sentiment Analysis Using CBGA Hybrid Deep Learning Model[10]	Convolutional Bidirectional Gated Recurrent Unit (Bi-GRU) and CBGA	Measure the performance for the proposed model using precision, sensitivity, accuracy, and Matthews Correlation Coefficient (MCC).
10	Social Media Mining: a Genetic-Based Multiobjective Clustering Approach to Topic Modelling[11]	Multiobjective Genetic algorithm (MOGA)	Including clustering techniques and applied Genetic algorithms using Reproduction crossover and mutation.
11	LibShortText: a Library for Short-text Classification and Analysis[12]	The research did not include LDA or LSI	Implementing a new tool (LibShortText) to classify the short text and support the research by algorithms that are integrated
12	Comparison of Topic Modelling Approaches in the Banking Context[22]	LDA is not applied in this research	BERT topic architecture using Kernal Principal Component Analysis (PCA) and K-Means to identify the coherent topics

Furthermore, Wang, W. et al. used text mining and Bayesian inference in their study to identify and differentiate the subjective and objective attributes in an entrepreneurial narrative in crowdfunding projects. Also, he published another research to produce different text topics during different stages -antecedent stage and posterior stage-. Based on the comment text of online financing projects, this study analyses the dynamic topics evolution in the online comments, considering both the time and topic factors.(13,14) Peng, N. et al.(15) took the advantage of coronavirus pandemic and researched medical crowdfunding to predict the fundraising rate and promote sustainable development in medical crowdfunding. Although there is a lot of research focusing on text classification to classify different types of documents, there is rare text classification research that used crowdfunding data to apply the classification in terms of topic categorization, and that what is characterizes this research.(16,17,18) The necessity of creating new projects, whereas fund is the issue, hence this research would create awareness on crowdfunding and applying the IT machine learning techniques which would combine innovation and finances to support the socio-economic needs especially supporting the SMEs and start-ups.(19,20) In turn, using crowdfunding data will help the funders to raise the money by preparing ready simple data about the available projects. Finally, this research also will raise awareness about crowdfunding among the people.

METHOD

Topic modeling aims to find the pattern that frequently reflects the fundamental topics that are joined to predict the topics that are derived from documents. For example, hierarchical probabilistic models effortlessly classified different sorts of data; TM has been utilized to evaluate words inside any documents and classify the text depending on the categories. The fundamental significance of TM is to figure out a pattern of words and how to interface documents that dispense the comparative pattern. So, the goal of TM is working with documents and these documents are combinations of topics, where a subject is a likelihood conveyance over words. So, TM is considered a generative model for documents. It determines a basic probabilistic strategy by which documents can be created, then builds another document by picking a conveyance over topics. From that point onward, each word in the document could pick a topic blindly rely upon the appropriation, and lastly construct a word from that point. So, this chapter will discuss the development approach and present the selected framework, which includes the tools and the dataset, the pre-processing procedures, applying the algorithms ending by measuring the performance. Natural language Processing (NLP) is a demanding study in computer science to analyze the text, semantic mining, and empowering PCs to predict the explanation from human language allocated in Text content. TM strategies are sufficient procedures that are broadly applied in NLP to point revelation and semantic mining from non-organized documents and data. In a wider range, several methods are used in the TM field such as Latent Dirichlet Allocation (LDA), Parallel Latent Dirichlet Allocation (PLDA), Deterministic Finite Automata (DFA), Latent Semantic Analysis (LSA), etc. LSA and LDA are commonly used as Topic Modelling methods, where there are a lot of similarities. Both depend on bag-of-words modeling, starting by converting the text corpora to a term-document prevalence array. Then decrease the high dimensional term spaces of textual data to a client-identified number of dimensions provide weighted term records for each identification or topic, and present identification or topic context weights for each document. Finally deliver outputs are adapted to compute document association evaluations. However, regardless of these similitudes, the two algorithms LSA and LDA produce different outputs. LSA utilizes singular value decomposition (SVD) to characterize a reason to a combined semantic vector space, in which the most extreme difference beyond the information is caught in a specific number of dimensions. Conversely, LDA utilizes a Bayesian model to deal with every record as a combination of latent elemental topics, where every topic is displayed as a combination of word possibilities from a vocabulary. Even though LSA and LDA outputs can be utilized likewise, their outputs show entirely various quantities, with various parameters and concepts. LSA generates term concept and archive-concept connection matrixes, with values going somewhere in the range of −1 and 1 with negative qualities demonstrating reverse relationships. Figure 4 represents various classification algorithms.

Text classification is divided into two categories machine learning and statistical. Machine learning takes a huge role in text classification and there are three divisions of it depending on the supervision. The first sector is supervised learning which is divided into parametric and non-parametric classifiers and there are methods branching from both. The second sector is semi-supervised learning which includes self-training, transductive SV, etc. The last sector is unsupervised machine learning which contains k-means clustering, hierarchal clustering, and fuzzy C-means. This classification algorithm of figure 4, which represents the input parameters of DSTD will be moved into the text segmentation, it holds the relationship between text-based documents and topics to select the optimal topics. For selecting the optimal topics, the LDA/LSA preprocessing techniques are utilized for text grouping and distribution of documents and words.

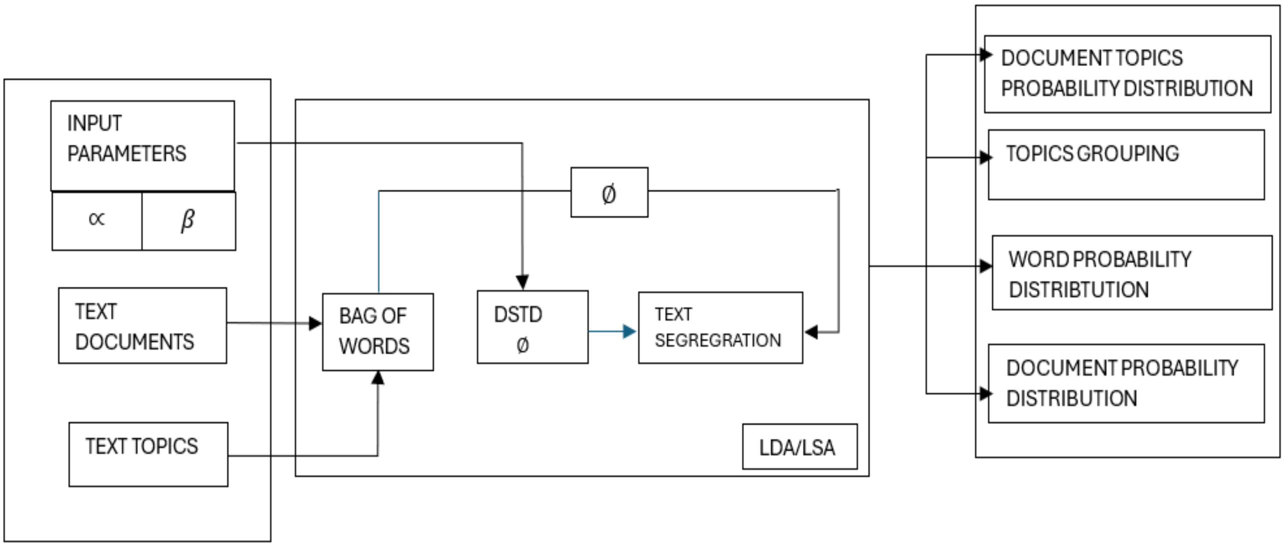
INPUT PARAMETERS
∝
β
TEXT DOCUMENTS
TEXT TOPICS
BAG OF WORDS
Ø
DSTD Ø
TEXT SEGREGRATION
LDA/LSA
DOCUMENT TOPICS PROBABILITY DISTRIBUTION
TOPICS GROUPING
WORD PROBABILITY DISTRIBTUTION
DOCUMENT PROBABILITY DISTRIBUTION

Figure 4. Classification algorithms

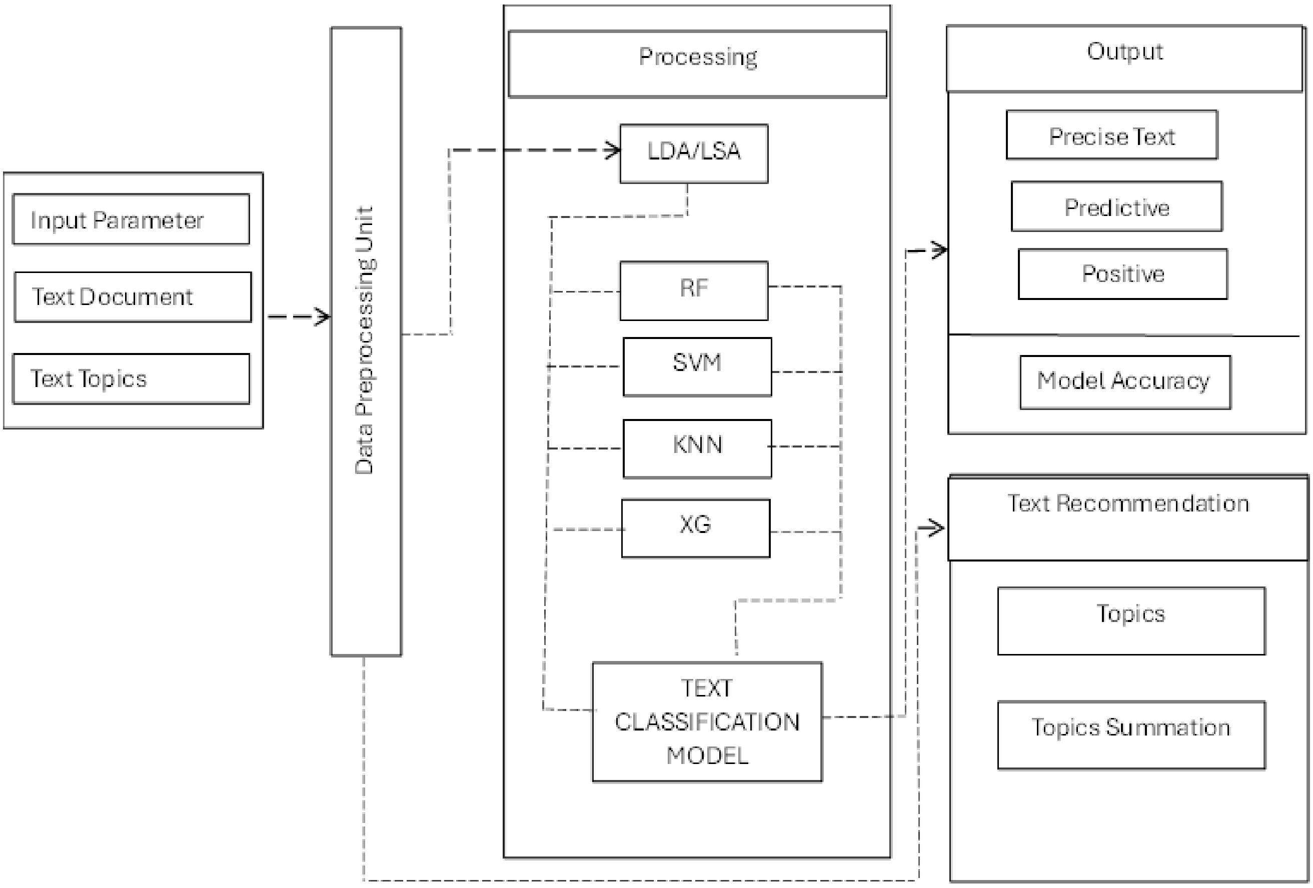

Figure 5. Text Framework of Optimization

Data Source

The information utilized in this experiment was obtained online from Kaggle.com.[23] It's a CSV file containing data on the Kickstarter project. It is made up of 378 054 subjects divided into 10 columns. The most essential aspects included in this research are the title, which comprises the topic name and the category of each topic. This material is organized into categories such as Art, Fashion, Dance, Academic, Places, Festivals, and Photography. Before cleaning, the CSV files contained non-English words, symbols, and numbers.

Proposed Algorithm

Algorithm 1: classification and recommendation

```
for text topics N = 1-N
select text topics $Ø_N$ ~ Dir(∝,β)
for text document $T_d = 1 \rightarrow T_D$
            KNN model initialization with $K_0 = 0$
            SVM model initialization with $S_0 = 0$
            RF model initialization with $R_0 = 0$
            XG model initialization with $X_0 = 0$
            for words at $T_t = 1 \rightarrow K$ in document $T_w$
 proposed model update as $T_k$ = model $(XT_w)$
 probability distribution = topics + document
 probability weight = topic + words
                        select a suitable topic from the document
                        select the appropriate word from the topics
        recommend the accuracy of the model
```

Table 2. LDA/LSA parameters and definitions

Parameters	Type	Definition
N	Int	No. Of. Topics
DV	Int	Word repository
T_d	Int	Collection of documents
T_w	Int	Collection of Word in a document
T_t	Int	Counting of words from T_w
alpha	Dim	Dimension measurement topics and words from a document
beta		
theta		
x	N-dim	Select the topic from the document

RESULTS

In this proposed implementation part of the system and findings. Latent District Allocation (LDA) and Latent Semantic Allocation (LSA) will be tested as topic modeling methods. As well as, Semantic Vector Machine (SVM), Extreme Gradient Boosting (XG), K-Nearest Neighbour (KNN), and Random Forest (RF) were integrated to classify the text. Then the performance is evaluated and compared among the algorithms depending on the accuracy, precision, Recall, and F1. At the end, the findings as presented and explained as well. The system will start first with the pre-processing stage to clean the data and prepare it for the next algorithms. So, it will start by reading the data, removing the nulls, changing the words to lowercase, and removing the special characters. Additionally, short words with less than 4 letters will be removed. As well as the non-English word will be excluded. As a result, empty rows will be excluded so the system will cancel the empty rows. And last the name and blurb columns will be combined. Sample of the data after the pre-processing stage. Implementing LDA and LSA Methods: after the pre-processing stage several libraries were imported to use with LDA and LSA methods such as Sklearn.feature_extraction.text import TfidfVectorizer, Sklearn. decomposition import TruncatedSVD, Sklearn.feature_extraction.text import CountVectorizer, Sklearn. decomposition import LatentDirichletAllocation, Sklearn.pipeline, Numpy. After importing the required libraries, the system will apply the LSA and LDA in the CSV file that is used. 70 % of the CSV will be used to train the system while 30 % of the system will be used to test it. The system will generate a matrix in which rows are assigned to the documents and columns are assigned to the words. In LSA vectorizer is used and raw counts are replaced by tf-idfscore to assign a weight for the word appearance. Singular vector decomposition (SVD) is integrated in LSA to reduce the dimensionality by only selecting the largest singular value. On the other hand, in LDA each topic will share a common Dirichlet prior and be represented as a probabilistic distribution over a group of words. It utilizes the corpus that includes the document, and the document includes the word to choose the multinomial distribution for the topic and the document.

Applying the Text Classification Methods

After extracting the features from LDA and LSA the text classification methods are applied to classify the text and predict the category. So, in each algorithm, SVM, KNN, XG, and RF the system is trained using 70 % of the data and testing the remaining data. So, the system will generate a CSV file for each classifier that includes the category for each topic, it shows an example of category prediction in XG.

DISCUSSION

To measure the performance of each algorithm the researcher used four performance measurements:

Precise: which measures the accuracy of the classifier.

Predictive: it is the positive predictive value that measures the consistency of the results. Positive rate: it is called also the sensitivity to specify the fraction of relevant instances. Model accuracy: it is the weighted average of the predictive and positive rate measures.

Semantic Vector Machine Algorithm recorded the highest performance among the other classifiers, the precise was 91,2 %. On the other hand, the performance in the K-nearest neighbor was the lowest in accuracy at 62,30 %. extreme programming fell in between SVM and KNN in terms of accuracy at 82,20 %, while random forest came after SV at 91,10 %. The illustrates the precision measurement percentage for each text classification method that is used in this research. the Semantic Vector Machine Algorithm (SVM) also scored the highest precision percentage among the other classifiers. It was 88 % when applied to the dataset that is used in this research. RF algorithm came next after SVM at 93 %. The lowest precision percentage was registered for the KNN algorithm at 63,2 % close to the model accuracy measurement.

The positive rate measurement measures the number of correct positive results divided into positive results that must be returned. For SVM, KNN, XG, and RF methods recall was almost equal to the accuracy percentage for each one of them. Starting from the lowest the KNN was 69,5 % XG was 82,2 % to the highest at 94,4 % for SVM and RF at 84 %.

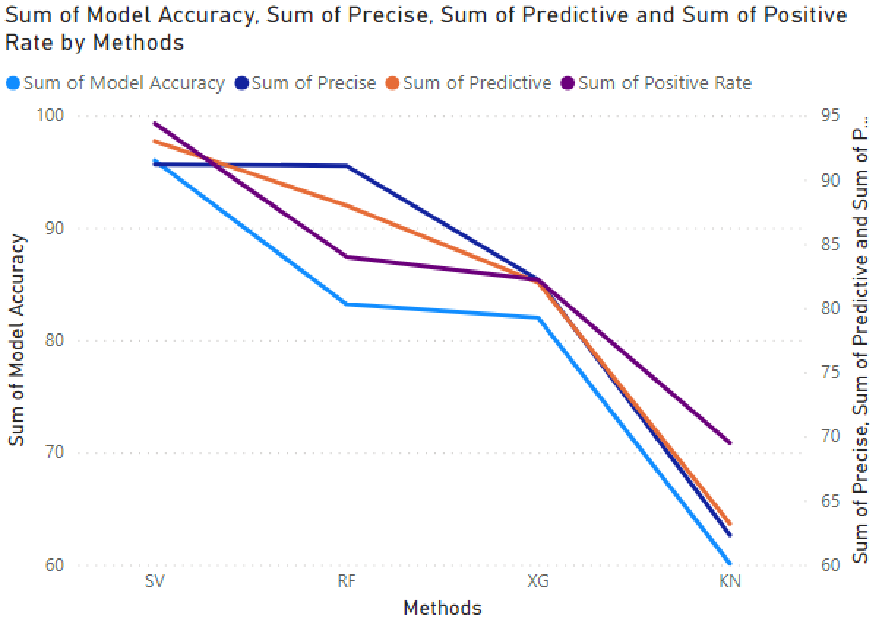

Figure 6. Model accuracy for failure rate

The model accuracy measurement weighs the average of the positive rate and the precision of the classifiers. As mentioned the average of the model measurement for the dataset that is used is between 60 % and 96 % for SVM and KNN. The KNN is the lowest at 60,1 % and the SVM is the highest at 96 % compared between each classifier in terms of the performance measurement.

Table 3. Proposed model techniques

Methods	Precise	Predictive	Positive Rate	Model Accuracy
SVM	91,20 %	93 %	94,4 %	96 %
RF	91,1 %	88 %	84 %	83,2 %
KNN	62,30 %	63,2 %	69,5 %	60,1 %
XG	82,2 %	82 %	82,2 %	82

The research merged the category information based on the existing LDA and LSA feature selection algorithm to explore the difference of underlying topics among the disparate class documents, and the simplified multi-class textual data have been sorted using SVM, KNN, XG, and RF classifiers. LDA and LSA have peerless modeling strengths, while the text classification algorithms have incomparable excellent properties on text categorization. So, the good text representation performance of the former with the powerful classification ability of the latter were combined to be more beneficial Part of the system is recommending topics for crowdfunding projects based on unstructured data, So the LDA model was used to predict the topic as shown below. In the LDA model, the topic prediction will be presented as a group of words extracted from the dictionary for each topic. So, the words will be sequenced after each other its samples of LDA prediction.

After accomplishing the experiment using the SVM, RF, KNN, and XG classifiers, the results show that the accuracy differs for each algorithm. The SVM recorded the highest accuracy score at 96 % then RF at 91,1 % while the XG is the lowest at 82,2 %. The precise scores almost equaled the recall score for each algorithm RF, SVM, and KNN. As well as SVM has the highest precision score at 91,20 % while the KNN is the lowest at 60,1 %. Finally, SVM has the highest model accuracy score while KNN is the lowest. So, depending on the previous findings and results SVM is the best text classifier for this dataset. Also, the research included topic prediction using the LDA model. So, it will start with categorizing the topic into specific categories depending on the description that is submitted on the data, then it will propose a topic for each row. Additionally, the research will raise awareness about crowdfunding and how it's related to topic modeling and machine learning by publishing the research on global websites and providing infographics that will be published using social media.

CONCLUSIONS

Crowdfunding has risen as the need for finance and business has risen in parallel. It is a sufficient method to create capital through the support of customers, and investors. It is like a large pool of projects that need to be adopted and supported. This pool is presented as online social media and crowdfunding platforms. It has gained a reputation and attention from SMEs, investors, policymakers, and researchers. This research combined Latent Dirichlet Allocation and Latent Semantic Analysis in parallel with Semantic Vector Machine (SVM), Extreme Gradient Boosting (XG), K Nearest Neighbour (KNN), and Random Forest (RF). While most studies used them separately or some of them not most. Also so far, few researchers started to classify the crowdfunding data but it's rare to find research that implemented text classification methods and merged machine learning in crowdfunding fields so it will contribute to helping the investors in crowdfunding projects. Additionally, rare research classified crowdfunding data in terms of topic/ project category and proposed new topic. Most of them depend on customers' reviews, the rate of funds, and the success of the project. This research contributes to raising awareness about crowdfunding by publishing the research after it's approved on global websites. As well as designing an infographic brochure that includes information about crowdfunding, its platforms, and the benefits of publishing it on social media.

BIBLIOGRAPHIC REFERENCES

1. Jin W. Research on Machine Learning and Its Algorithms and Development. J Phys Conf Ser. 2020;1544(1). doi: 10.1088/1742-6596/1544/1/012003.

2. Kolluri J, Razia S, Nayak SR. Text Classification Using Machine Learning and Deep Learning Models. SSRN Electron J. 2020. doi: 10.2139/ssrn.3618895.

3. Aljebory KM, Jwmah YM, Mohammed TS. Classification of EMG Signals: Using DWT Features and ANN Classifier. IAENG Int J Comput Sci. 2024;51(1).

4. Naqa IE, Murphy MJ. Machine Learning in Radiation Oncology. In: Machine Learning in Radiation Oncology. 2015. p. 3-11. doi: 10.1007/978-3-319-18305-3.

5. Hannigan TR, et al. Topic modelling in management research: Rendering new theory from textual data. Acad Manag Ann. 2019;13(2):586-632. doi: 10.5465/annals.2017.0099.

6. Hasan M, Hossain MM, Ahmed A, Rahman MS. Topic Modelling: A Comparison of the Performance of Latent Dirichlet Allocation and LDA2vec Model on Bangla Newspaper. 2019 Int Conf Bangla Speech Lang Process ICBSLP. 2019 Sep;27-8. doi: 10.1109/ICBSLP47725.2019.202047.

7. Shafqat W, Byun YC. A recommendation mechanism for under-emphasized tourist spots using topic modelling and sentiment analysis. Sustainability. 2020;12(1). doi: 10.3390/SU12010320.

8. Debortoli S, Müller O, Junglas I, vom Brocke J. Text Mining For Information Systems Researchers: An Annotated Topic Modeling Tutorial. Commun Assoc Inf Syst. 2016;39. doi: 10.17705/1CAIS.03907.

9. Iyyappan M, Ahmad S, Jha S, Alam A, Yaseen M, Hikmat A. A Novel AI-Based Stock Market Prediction Using Machine Learning Algorithm. Sci Program. 2022;1-11. doi: 10.1155/2022/4808088.

10. Wei L, et al. A Lightweight Sentiment Analysis Framework for a Micro-Intelligent Terminal. Sensors. 2023;23(2):741. doi: 10.3390/s23020741.

11. Alfred R, Loo YJ, Obit JH, Lim Y, Haviluddin H, Azman A. Social media mining: a genetically based multiobjective clustering approach to topic modelling. 2021.

12. Hsiang FY, Chia HH, Yu CJ, Chih JL. LibShortText: A Library for Short-text Classification and Analysis. 2013. Available at: https://www.csie.ntu.edu.tw/~cjlin/libshorttext/.

13. Wang W, et al. Signaling persuasion in crowdfunding entrepreneurial narratives: The subjectivity vs objectivity debate. Comput Human Behav. 2021;114(Sep 2020):106576. doi: 10.1016/j.chb.2020.106576.

14. Wang W, Wu YJ. Online Financing Campaigns' Comments: Insights from Crowdfunding Pitches. In: Visvizi A, Lytras MD, Aljohani NR, editors. Research and Innovation Forum 2020 - Disruptive Technologies in Times of Change. Springer; 2021. p. 485-93. doi: 10.1007/978-3-030-62066-0_37.

15. Peng N, et al. Predicting fundraising performance in medical crowdfunding campaigns using machine learning. Electronics (Switzerland). 2021;10(2):1-16. doi: 10.3390/electronics10020143.

16. Yuan H, Lau RYK, Xu W. The determinants of crowdfunding success: A semantic text analytics approach. Decis Support Syst. 2016;91:67-76. doi: 10.1016/j.dss.2016.08.001.

17. Zhao Y, Harris P, Lam W. Crowdfunding industry—History, development, policies, and potential issues. J Public Aff. 2019;19(1). doi: 10.1002/pa.1921.

18. Robinson D. broom: An R Package for Converting Statistical Analysis Objects Into Tidy Data Frames. 2014. Available at: https://arxiv.org/abs/1412.3565.

19. Slavik S. The Business Model of Start-Up — Structure and Consequences. Adm Sci. 2019;9(69):1-23. doi: 10.3390/admsci9030069.

20. McGowan E. What is Crowdfunding? 2018. Available at: https://www.startups.com/library/expert-advice/what-is-crowdfunding.

21. Uthirapathy SE, Sandanam D. Topic Modelling and Opinion Analysis On Climate Change Twitter Data Using LDA And BERT Model. Procedia Comput Sci. 2023;218:908-17.

22. Ogunleye B, Maswera T, Hirsch L, Gaudoin J, Brunsdon T. Comparison of topic modelling approaches in the banking context. Appl Sci. 2023;13(2):797.

23. N R, Nachiappan B, Kalpana C, Mohanraj A, Prabhu Shankar B, Viji C. Machine Learning-Based System for Automated Presentation Generation from CSV Data. Data and Metadata [Internet]. 2024 Jul. 2 [cited 2024 Jul. 8];3:359. Available from: https://dm.saludcyt.ar/index.php/dm/article/view/359.

24. Nachiappan B, Rajkumar N, Viji C, A M. Artificial and Deceitful Faces Detection Using Machine Learning. Salud, Ciencia y Tecnología - Serie de Conferencias [Internet]. 2024 Mar. 11 [cited 2024 Jul. 8];3:611. Available from: https://conferencias.saludcyt.ar/index.php/sctconf/article/view/611

FINANCING

No financing.

CONFLICT OF INTEREST

None.

AUTHORSHIP CONTRIBUTION

Conceptualization: Suresh Subramanian.
Data curation: Suresh Subramanian.
Formal analysis: Suresh Subramanian.
Drafting - original draft: Suresh Subramanian.
Writing - proofreading and editing: Suresh Subramanian.